

Optimal decision for the market graph identification problem in sign similarity network.

Kalyagin V.A., Koldanov P.A., Pardalos P.M.¹

Abstract: Investigation of the market graph attracts a growing attention in market network analysis. One of the important problem connected with market graph is to identify it from observations. Traditional way for the market graph identification is to use a simple procedure based on statistical estimations of Pearson correlations between pairs of stocks. Recently a new class of statistical procedures for the market graph identification was introduced and optimality of these procedures in Pearson correlation Gaussian network was proved. However the obtained procedures have a high reliability only for Gaussian multivariate distributions of stocks attributes. One of the way to correct this drawback is to consider a different networks generated by different measures of pairwise similarity of stocks. A new and promising model in this context is the sign similarity network. In the present paper the market graph identification problem in sign similarity network is considered. A new class of statistical procedures for the market graph identification is introduced and optimality of these procedures is proved. Numerical experiments detect essential difference in quality of optimal procedures in sign similarity and Pearson correlation networks. In particular it is observed that the quality of optimal identification procedure in sign similarity network is not sensitive to the assumptions on distribution of stocks attributes.

Keywords: Pearson correlation network, sign similarity network, market graph, multiple decision statistical procedures, loss function, risk function, optimal multiple decision procedures.

1 Introduction

There are a variety of data mining techniques applied for stock markets. Some of them are based on the analysis of market network and its structures. Market network is a complete weighted graph where the nodes are associated with stocks and weights of edges are given by some measure of similarity between stocks behavior. *Market graph* is an important structure in market network. An edge between two nodes is included in the market graph, iff the corresponding measure of similarity is larger than a given threshold. Maximum cliques, maximum independent sets, degree distribution in the market graph are useful sources of market data mining.

The concept of the market graph was introduced in [3]. Since, different aspects of the market graph approach (threshold method) were developed in the literature. Most publications are related with experimental study of real markets. The power law phenomena first observed for US stock market in

¹National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, Bolshaya Pecherskaya 25/12, Nizhny Novgorod, RUSSIA. Tel.: +7-831-4361397. Fax: +7-831-4169655. Email: vkalyagin@hse.ru

[4] was then developed in [10], [22], [6]. Clustering in Pearson correlation based financial network is investigated in [20]. Dynamics of the US market graphs was studied in [5]. Complexity of the US market graph associated with significant correlations is investigated in [7]. Peculiarity of different financial markets are emphasized in [1], [8], [24], [10], [19]. Market graphs with different measures of similarity were investigated in [2], [1], [11], [21], [26], [13]. Some efficient algorithms related with calculation of isolated cliques in a market graph are presented in [9], [12].

However, economical interpretation of market network data mining is not complete without estimation of reliability of obtained results. Reliability of minimum spanning tree in Pearson correlation based network is investigated by bootstrap method in [23]. In the present paper we use a different approach to handle reliability of the market graph. Our approach is based on the model of *random variables network*. The nodes of the network are random variables and weights of edges are given by some measure of pairwise similarity between the random variables. Observed values of stocks attributes are modeled by sample from distribution of random variables. Reliability of network structure can now be measured by risk function of statistical procedure for its identification. Statistical procedures with minimal risk (maximal reliability) are of great practical interest. Class of optimal statistical procedures for the identification of market graph in Pearson correlation based network were introduced and investigated in [14]. One can see from numerical experiments in [14] that the value of the risk function of the procedures of this class essentially depends on the assumptions on multivariate distributions of stocks attributes. Taking this into account it is of interest to investigate a distribution free identification statistical procedures. A new and promising approach in this context is to consider a sign correlation based (sign similarity) network.

In this paper we investigate optimal statistical procedures for the market graph identification in sign similarity (sign correlation based) network. Our construction of optimal procedures is based on simultaneous inference of optimal two-decision procedures. It is proved that constructed procedure is optimal under the following assumptions: additivity of loss function, unbiasedness of procedure, sign symmetry of distributions. We give a direct proof which simplifies a general approach by Lehmann [16]. In addition we compare the risk function of the optimal procedure in sign similarity network with the risk function of the optimal procedure in Pearson correlation network. Numerical experiments detect essential difference in risk behavior for two optimal statistical procedures. For multivariate Gaussian distribution both procedures control the risk with a change of significance level of individual tests. In contrast for multivariate Student distribution optimal procedure in Pearson correlation network does not control the risk while the optimal procedure in sign similarity network does. It means that the quality of optimal identification procedure in sign similarity network is not sensitive to the assumptions on distribution of stocks attributes.

The paper is organized as follows. In the section 2 we give a basic definitions and notations. In section 3 we describe a multiple decision framework for threshold graph identification problem. In Section 4 we discuss a concept

of optimality of identification procedures. In Section 5 we construct a multiple decision identification procedure in sign similarity network. In Section 6 we give a proof of optimality of this procedure. In section 7 we conduct a numerical experiments to compare optimal procedures in sign similarity and Pearson correlation network. In section 8 we present a concluding remarks.

2 Market graph identification problem

Consider a network generated by a random vector $X = (X_1, X_2, \dots, X_N)$. Nodes of network are random variables X_i , $i = 1, \dots, N$ and weight of edge (i, j) is given by some pairwise measure of association γ :

$$\gamma_{i,j} = \gamma(X_i, X_j), \text{ for } i, j = 1, 2, \dots, N.$$

The obtained network is a complete weighted graph which we will call *random variables network*. The random variables network is defined by multivariate distribution of the vector X and by the choice of measure of association γ . The network based on Pearson correlation of random variables will be called *Pearson correlation network*. The network based on probability of pairwise sign coincidence will be called *sign similarity network*.

For any network the market graph is constructed as follows: the edge between two vertices i and j is included in the market graph, iff $\gamma_{i,j} > \gamma_0$, where γ_0 is a given threshold. In what follows we will call this network structure *reference (true) market graph*.

In sign similarity network weight of edge (i, j) is defined by

$$p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j)) > 0) \quad (1)$$

For a given threshold p_0 *reference market graph* in sign similarity network is constructed as follows: edge between two nodes i and j is included in the reference market graph iff $p^{i,j} > p_0$, where $p^{i,j}$ is the probability of sign coincidence of random variables associated with nodes i and j .

In practice $\gamma_{i,j}$ are unknown and we are given a sample of observations $x(1), x(2), \dots, x(t)$ from distribution X . Identification of reference market graph from observations is called in this paper *market graph identification problem*.

Two types of errors are possible for any identification procedure. Type I error occurs if identification procedures includes edge in the market graph when it is absent in the reference market graph. Type II error occurs if identification procedures does not include edge in the market graph when it is present in the reference market graph. For market graph identification it is important to control not only type I and type II errors but the number of errors.

3 Multiple decision framework

We model observations as a family of random vectors

$$X(t) = (X_1(t), X_2(t), \dots, X_N(t)), \quad t = 1, 2, \dots, n$$

where n is the number of observations (sample size) and vectors $X(t)$ are independent and identically distributed as $X = (X_1, X_2, \dots, X_N)$. In what follows we assume that expectations $E(X_i)$, $i = 1, 2, \dots, N$ are known. We put (for simplicity) $E(X_i) = 0$, $i = 1, 2, \dots, N$. In this case

$$p^{i,j} = P(X_i X_j > 0), \quad i, j = 1, 2, \dots, N \quad (2)$$

The random vector X with measures of association $(p^{i,j})$ define a sign similarity network. For a given threshold p_0 the reference market graph is defined by its adjacency matrix $TG = (tg_{i,j})$, where $tg_{i,j} = 0$ if $p^{i,j} \leq p_0$ and $tg_{i,j} = 1$ if $p^{i,j} > p_0$, $tg_{i,i} = 0$, $i, j = 1, 2, \dots, N$.

Let $x_i(t)$ be observations of the random variables $X_i(t)$, $t = 1, 2, \dots, n$, $i = 1, 2, \dots, N$. Consider the set \mathcal{G} of all $N \times N$ symmetric matrices $G = (g_{i,j})$ with $g_{i,j} \in \{0, 1\}$, $i, j = 1, 2, \dots, N$, $g_{i,i} = 0$, $i = 1, 2, \dots, N$. Matrices $G \in \mathcal{G}$ represent adjacency matrices of all simple undirected graphs with N vertices. The total number of matrices in \mathcal{G} is equal to $L = 2^M$ with $M = N(N - 1)/2$. The *market graph identification problem* in sign similarity network can be formulated as a multiple decision problem of the selection of one from the set of L hypotheses:

$$H_G : p^{i,j} \leq p_0, \text{ if } g_{i,j} = 0, \quad p^{i,j} > p_0, \text{ if } g_{i,j} = 1; \quad i \neq j \quad (3)$$

Consider some examples of hypotheses (3). For the matrix

$$G_1 = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

the corresponding graph has N isolated vertices and the associated hypothesis H_{G_1} is

$$H_{G_1} : p^{i,j} \leq p_0, \forall i, j = 1, \dots, N, \quad i \neq j$$

For the matrix:

$$G_2 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

the corresponding graph has only one edge $(1, 2)$ and the associated hypothesis H_{G_2} is

$$H_{G_2} : p^{1,2} > p_0, p^{2,1} > p_0, p^{i,j} \leq p_0, \forall (i, j) \neq (1, 2), (i, j) \neq (2, 1), i \neq j$$

For the matrix:

$$G_3 = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 0 \end{pmatrix}.$$

the corresponding graph is a complete graph and the associated hypothesis H_{G_3} is

$$H_{G_3} : p^{i,j} > p_0, \forall i, j = 1, \dots, N, \quad i \neq j$$

To solve the identification problem (3) one has to use multiple decision statistical procedures. Multiple decision statistical procedure δ is a map from the sample space $R^{N \times n}$ to the decision space $D = \{d_G, g \in \mathcal{G}\}$, where the decision d_G is the acceptance of hypothesis H_G , $G \in \mathcal{G}$.

4 Concept of optimality

In this section we discuss a concept of optimality related with multiple decision statistical procedures. According to [25] the quality of statistical procedure is defined by risk function. Consider a statistical procedure $\delta(x)$. Let $S = (s_{i,j})$, $Q = (q_{i,j})$, $S, Q \in \mathcal{G}$. Denote by $w(S, Q)$ the loss from the decision d_Q when the hypothesis H_S is true

$$w(H_S; d_Q) = w(S, Q), \quad S, Q \in \mathcal{G}$$

It is assumed that $w(S, S) = 0$, $S \in \mathcal{G}$. Risk function $Risk : \mathcal{G} \rightarrow \mathcal{R}$ is defined by

$$Risk(S; \delta) = \sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q / H_S), \quad S \in \mathcal{G}$$

where $P(\delta(x) = d_Q / H_S)$ is the probability that decision d_Q is taken while the true decision is d_S . The problem of minimization of $Risk$ is therefore multiple criteria decision problem: optimal procedure δ has to minimize $Risk(S, \delta)$ for every $S \in \mathcal{G}$. In general such problem does not have a solution. Instead one can use a Pareto optimal solutions. However, it is possible to get a solution if one imposes a constraints on the procedures. One common constraint is unbiasedness of the procedure. Following [17] we call the multiple decision procedure $\delta(x)$ w -unbiased if

$$\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q / H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta(x) = d_Q / H_S), \quad \forall S, S' \in \mathcal{G} \quad (4)$$

“Thus δ is unbiased if on the average $\delta(x)$ comes closer to the correct decision than to any wrong one” (citation from [17], page 13). Note that unbiasedness depends on the loss function w .

For the market graph identification problem it is important to control not only type I and type II errors but a number of errors. Let $a_{i,j}$ be the loss from false inclusion of edge (i, j) in market graph and let $b_{i,j}$ be the loss from false non inclusion of the edge (i, j) in the market graph, $i, j = 1, 2, \dots, N$; $i \neq j$.

Let

$$l_{i,j}(S, Q) = \begin{cases} a_{i,j}, & \text{if } s_{i,j} = 0, q_{i,j} = 1, \\ b_{i,j}, & \text{if } s_{i,j} = 1, q_{i,j} = 0, \\ 0, & \text{else} \end{cases}$$

It is natural to define loss function $w(S, Q)$ as:

$$w(S, Q) = \sum_{i=1}^N \sum_{j=1}^N l_{i,j}(S, Q) \quad (5)$$

It means that the loss from misclassification of H_S is equal to the sum of losses from misclassification of individual edges:

$$w(S, Q) = \sum_{\{i,j:s_{i,j}=0;q_{i,j}=1\}} a_{i,j} + \sum_{\{i,j:s_{i,j}=1;q_{i,j}=0\}} b_{i,j}$$

In the next sections we investigate the optimality of multiple statistical procedures for market graph identification for the loss function in the class of unbiased procedures.

5 Multiple decision procedure based on simultaneous inference of two decision tests

In this section we describe a class of multiple decision procedures based on simultaneous inference of individual edge tests.

Any individual edge test can be reduced to hypotheses testing problem: $h_{ij} : \gamma_{ij} \leq \gamma_0$ vs $k_{ij} : \gamma_{ij} > \gamma_0$. According to [17] test of the individual edge hypotheses has the form:

$$\varphi_{ij}(x) = \begin{cases} 0, & t_{ij}(x) \leq c_\alpha \\ 1, & t_{ij}(x) > c_\alpha \end{cases} \quad (6)$$

where $\varphi_{ij}(x) = 1$ means that edge (i, j) is included in the market graph, $\varphi_{ij}(x) = 0$ means that edge (i, j) does not included in the market graph.

Let $\Phi(x)$ be the matrix

$$\Phi(x) = \begin{pmatrix} 1, & \varphi_{12}(x), & \dots, & \varphi_{1N}(x) \\ \varphi_{21}(x), & 1, & \dots, & \varphi_{2N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}(x), & \varphi_{N2}(x), & \dots, & 1 \end{pmatrix}. \quad (7)$$

Then all statistical procedures from the class of multiple decision procedures based on simultaneous inference of individual edge tests can be written as

$$\delta(x) = d_G, \text{ iff } \Phi(x) = G \quad (8)$$

Consider the sign similarity network. Let

$$\begin{aligned} p_{0,0}^{i,j} &= P(X_i \leq 0, X_j \leq 0), & p_{1,1}^{i,j} &= P(X_i > 0, X_j > 0) \\ p_{1,0}^{i,j} &= P(X_i > 0, X_j \leq 0), & p_{0,1}^{i,j} &= P(X_i \leq 0, X_j > 0) \end{aligned}$$

One has $p^{i,j} = p_{0,0}^{i,j} + p_{1,1}^{i,j}$. Define

$$u_k(t) = \begin{cases} 0, & x_k(t) \leq 0 \\ 1, & x_k(t) > 0 \end{cases}$$

$k = 1, 2, \dots, N$. Let us introduce statistics

$$\begin{aligned} T_{1,1}^{i,j} &= \sum_{t=1}^n u_i(t)u_j(t); & T_{0,0}^{i,j} &= \sum_{t=1}^n (1-u_i(t))(1-u_j(t)); \\ T_{0,1}^{i,j} &= \sum_{t=1}^n (1-u_i(t))u_j(t); & T_{1,0}^{i,j} &= \sum_{t=1}^n u_i(t)(1-u_j(t)); \\ V_{i,j} &= T_{1,1}^{i,j} + T_{0,0}^{i,j} \end{aligned} \quad (9)$$

To construct a multiple decision procedure we use the following individual edge tests:

$$\varphi_{i,j}^{Sg}(x_i, x_j) = \begin{cases} 0, & V_{i,j} \leq c_{i,j} \\ 1, & V_{i,j} > c_{i,j} \end{cases} \quad (10)$$

where for a given significance level $\alpha_{i,j}$, the constant $c_{i,j}$ is defined as minimal entire number such that:

$$\sum_{k=c_{i,j}}^n \frac{n!}{k!(n-k)!} (p_0)^k (1-p_0)^{n-k} \leq \alpha_{i,j} \quad (11)$$

Let $\Phi^{Sg}(x)$ be the matrix

$$\Phi^{Sg}(x) = \begin{pmatrix} 1, & \varphi_{12}^{Sg}(x), & \dots, & \varphi_{1N}^{Sg}(x) \\ \varphi_{21}^{Sg}(x), & 1, & \dots, & \varphi_{2N}^{Sg}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}^{Sg}(x), & \varphi_{N2}^{Sg}(x), & \dots, & 1 \end{pmatrix}. \quad (12)$$

where $\varphi_{ij}^{Sg}(x)$ are defined by (10)-(11). Now we can define our multiple decision statistical procedure for market graph identification

$$\delta^{Sg}(x) = d_G, \text{ iff } \Phi^{Sg}(x) = G \quad (13)$$

Constructed procedure looks very natural for market graph identification in sign similarity network.

6 Optimality of decision procedure δ^{Sg} .

In this section conditions of optimality of multiple decision statistical procedure defined by (10)-(13) among multiple decision statistical procedures based on two decision tests are described.

Theorem 1 *Let loss function w be given by (5), individual test statistics t_{ij} (6) depends only on $u_i(t), u_j(t)$, and following symmetry conditions are satisfied*

$$p_{11}^{ij} = p_{00}^{ij}, \quad p_{10}^{ij} = p_{01}^{ij}, \quad \forall i, j = 1, 2, \dots, N \quad (14)$$

Then for the statistical procedure δ^{Sg} defined by (10)-(13) for market graph identification in sign similarity network one has $Risk(S, \delta^{Sg}) \leq Risk(S, \delta)$ for any adjacency matrix S and any w -unbiased statistical procedure δ .

Proof We prove optimality in three steps. First we prove that under symmetry conditions (14) each individual test (10) is uniformly most powerful (UMP) in the class of tests based on $u_i(t), u_j(t)$ only for individual hypothesis testing

$$h_{i,j} : p^{i,j} \leq p_0 \quad \text{vs} \quad k_{i,j} : p^{i,j} > p_0 \quad (15)$$

By symmetry conditions individual hypothesis (15) can be written as:

$$h_{i,j} : p_{00}^{i,j} \leq \frac{p_0}{2} \quad \text{vs} \quad k_{i,j} : p_{00}^{i,j} > \frac{p_0}{2} \quad (16)$$

Let $p_{0,0} = p_{0,0}^{i,j}$; $p_{0,1} = p_{0,1}^{i,j}$; $p_{1,0} = p_{1,0}^{i,j}$; $p_{1,1} = p_{1,1}^{i,j}$; $T_{0,0} = T_{0,0}^{i,j}$; $T_{0,1} = T_{0,1}^{i,j}$; $T_{1,0} = T_{1,0}^{i,j}$; $T_{1,1} = T_{1,1}^{i,j}$. One has

$$T_{1,1} + T_{1,0} + T_{0,1} + T_{0,0} = n;$$

Symmetry condition implies

$$p_{0,0} + p_{1,0} = \frac{1}{2}$$

Let $t_{1,1}, t_{1,0}, t_{0,1}, t_{0,0}$ be a non negative entire numbers with $t_{1,1} + t_{1,0} + t_{0,1} + t_{0,0} = n$ and $C = n!/(t_{1,1}!t_{1,0}!t_{0,1}!t_{0,0}!)$. One has

$$\begin{aligned} P(T_{1,1} = t_{1,1}; T_{1,0} = t_{1,0}; T_{0,1} = t_{0,1}; T_{0,0} = t_{0,0}) &= C p_{1,1}^{t_{1,1}} p_{1,0}^{t_{1,0}} p_{0,1}^{t_{0,1}} p_{0,0}^{t_{0,0}} = \\ &= C p_{0,0}^{t_{1,1}+t_{0,0}} p_{1,0}^{t_{1,0}+t_{0,1}} = C_1 \exp\{(t_{1,1} + t_{0,0}) \ln \frac{p_{0,0}}{1/2 - p_{0,0}}\} \end{aligned}$$

where $C_1 = C(1/2 - p_{0,0})^n$.

Then hypotheses (16) are equivalent to the hypotheses:

$$h'_{i,j} : \ln\left(\frac{p_{0,0}}{1/2 - p_{0,0}}\right) \leq \ln\left(\frac{p_0}{1 - p_0}\right) \quad \text{vs} \quad k'_{i,j} : \ln\left(\frac{p_{0,0}}{1/2 - p_{0,0}}\right) > \ln\left(\frac{p_0}{1 - p_0}\right) \quad (17)$$

For $p_{0,0} = p_0/2$ random variable $V = T_{1,1} + T_{0,0}$ has the binomial distribution $B(n, p_0)$. Therefore, critical value for the test (10) is defined from (11). According to ([17], Ch.3, corollary 3.4.1) the test (10) is uniformly most powerful (UMP) at the level $\alpha_{i,j}$ for hypothesis testing (17).

Second we prove that statistical procedure (13) is w -unbiased. For any two-decision test for hypothesis testing (15) the risk function can be written as:

$$Risk = R(s_{i,j}, \varphi_{i,j}) = \begin{cases} a_{i,j} P(\varphi_{i,j}(x) = 1/p^{i,j}), & \text{if } s_{i,j} = 0 (p^{i,j} \leq p_0) \\ b_{i,j} P(\varphi_{i,j}(x) = 0/p^{i,j}), & \text{if } s_{i,j} = 1 (p^{i,j} > p_0) \end{cases}$$

Note that by general principle UMP test (10) is w -unbiased ([17], Ch.4). More precisely, one has

$$\begin{aligned} a_{i,j} P(\varphi_{i,j}^{Sg}(x) = 1/p^{i,j}) &\leq b_{i,j} P(\varphi_{i,j}^{Sg}(x) = 0/p^{i,j}) \quad \text{if } p_{i,j} \leq p_0 \\ a_{i,j} P(\varphi_{i,j}^{Sg}(x) = 1/p^{i,j}) &\geq b_{i,j} P(\varphi_{i,j}^{Sg}(x) = 0/p^{i,j}), \quad \text{if } p_{i,j} > p_0 \end{aligned}$$

which is equivalent to

$$R(s_{i,j}, \varphi_{i,j}^{Sg}) \leq R(s'_{i,j}, \varphi_{i,j}^{Sg}), \forall s_{i,j}, s'_{i,j} \quad (18)$$

This relation implies

$$P(\varphi_{i,j}^{Sg}(x) = 1/p^{i,j} = p_0) = \alpha_{i,j} = \frac{b_{i,j}}{a_{i,j} + b_{i,j}}$$

It means that test $\varphi_{i,j}^{Sg}$ has the significance level $\alpha_{i,j} = b_{i,j}/(a_{i,j} + b_{i,j})$. For loss function (5) and any multiple decision statistical procedure δ one has

$$\begin{aligned} R(H_S, \delta) &= \sum_{Q \in \mathcal{G}} (\sum_{i,j: s_{i,j}=0; q_{i,j}=1} a_{i,j} + \sum_{i,j: s_{i,j}=1; q_{i,j}=0} b_{i,j}) P(x \in D_Q/H_S) = \\ &= \sum_{s_{i,j}=0} a_{i,j} P(\varphi_{i,j}(x) = 1)/H_S + \sum_{s_{i,j}=1} b_{i,j} P(\varphi_{i,j}(x) = 0)/H_S \end{aligned} \quad (19)$$

Therefore:

$$R(H_S, \delta) = \sum_{i=1}^N \sum_{j=1}^N R(s_{i,j}; \varphi_{i,j}) \quad (20)$$

From (18) one has

$$\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta^{Sg}(x) = d_Q/H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta^{Sg}(x) = d_Q/H_S), \quad \forall S, S' \in \mathcal{G} \quad (21)$$

This means that multiple testing statistical procedure δ^{Sg} is unbiased.

Third we prove that procedure (13) is optimal in the class of unbiased statistical procedures for market graph identification in sign similarity network. Let $\delta(x)$ be another unbiased statistical procedure for market graph identification in sign similarity network. Then $\delta(x)$ generates a partition of sample space $R^{N \times n}$ on L parts:

$$D_G = \{x \in R^{N \times n} : \delta(x) = G\}; \bigcup_{G \in \mathcal{G}} D_G = R^{N \times n}$$

Define

$$\frac{A_{i,j}}{\bar{A}_{i,j}} = \frac{\bigcup_{G: g_{i,j}(x)=0} D_G}{\bigcup_{G: g_{i,j}(x)=1} D_G} \quad (22)$$

and

$$\varphi_{i,j}(x) = \begin{cases} 0, & x \in A_{i,j} \\ 1, & x \notin A_{i,j} \end{cases} \quad (23)$$

Tests (23) are tests for individual hypotheses testing (15). Since procedure $\delta(x)$ is unbiased then one has

$$\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q/H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta(x) = d_Q/H_S), \quad \forall S, S' \in \mathcal{G}$$

Consider the hypotheses H_S and $H_{S'}$ which are different only in two components $s_{i,j} \neq s'_{i,j}$; $s_{j,i} \neq s'_{j,i}$. Taking into account the unbiasedness of procedure δ and

structure of the loss function (5) one has $R(s_{i,j}, \varphi_{i,j}) \leq R(s'_{i,j}, \varphi_{i,j})$. This means that two decision tests (23) are unbiased. Therefore

$$P(\varphi_{i,j} = 1/p_0) = \alpha_{i,j} = \frac{b_{i,j}}{a_{i,j} + b_{i,j}}.$$

Since we are restricted the tests based on $u_i(t), u_j(t)$ only and test $\varphi_{i,j}^{Sg}$ is UMP among tests of the class at the significance level $\alpha_{i,j}$ then for any test $\varphi_{i,j}$ based on $u_i(t), u_j(t)$ only one has:

$$R(s_{i,j}, \varphi_{i,j}^{Sg}) \leq R(s_{i,j}, \varphi_{i,j})$$

From (20) one has

$$R(H_S, \delta^{Sg}) \leq R(H_S, \delta)$$

for any adjacency matrix S . Optimality of multiple testing statistical procedure δ^{Sg} has been proved.

7 Numerical comparison of optimal procedures in sign similarity and Pearson correlation networks

In this section we compare the behavior of risk functions for two optimal multiple decision procedures: identification procedure in sign similarity network (10)-(13) and identification procedure in Pearson correlation network with Gaussian distribution. Optimal identification procedure in Pearson correlation network with Gaussian distribution was investigated in [14]. This procedure can be presented as follows. Define sample Pearson correlations

$$r_{ij} = \frac{\sum_t x_i(t)x_j(t)}{\sqrt{\sum_t x_i(t)^2 \sum_t x_j(t)^2}}$$

Let ρ_0 be the threshold. We use the following individual edge tests:

$$\varphi_{i,j}^P(x_i, x_j) = \begin{cases} 0, & z_{i,j} \leq c_{i,j} \\ 1, & z_{i,j} > c_{i,j} \end{cases} \quad (24)$$

where

$$z_{i,j} = \sqrt{n} \left(\frac{1}{2} \ln \left(\frac{1 + r_{i,j}}{1 - r_{i,j}} \right) - \frac{1}{2} \ln \left(\frac{1 + \rho_0}{1 - \rho_0} \right) \right)$$

$c_{i,j}$ is $(1 - \alpha_{i,j})$ -quantile of standard normal distribution $N(0, 1)$, $\alpha_{i,j}$ is the given significance level.

Let $\Phi^P(x)$ be the matrix

$$\Phi^P(x) = \begin{pmatrix} 1, & \varphi_{12}^P(x), & \dots, & \varphi_{1N}^P(x) \\ \varphi_{21}^P(x), & 1, & \dots, & \varphi_{2N}^P(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}^P(x), & \varphi_{N2}^P(x), & \dots, & 1 \end{pmatrix}. \quad (25)$$

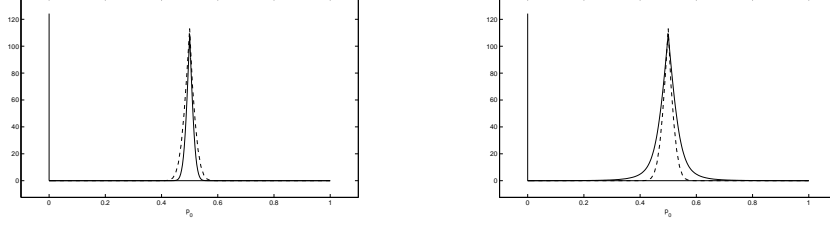


Figure 1: Risk functions $Risk(p_0)$ for the matrix Σ_1 and $\alpha = 0, 5$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axe represents the value of p_0 .

where $\varphi_{ij}^P(x)$ are defined by (24). Define the following multiple statistical procedure

$$\delta^P(x) = d_G, \text{ iff } \Phi^P(x) = G \quad (26)$$

It is proved in [14] that δ^P is optimal in Pearson correlation Gaussian network in the class of w -unbiased procedures, which have individual edge (i, j) tests based on observations $x_i(t), x_j(t) : i, j = 1, \dots, N$. Note, that both procedures δ^{Sg} and δ^P can be used for any distribution of vector X .

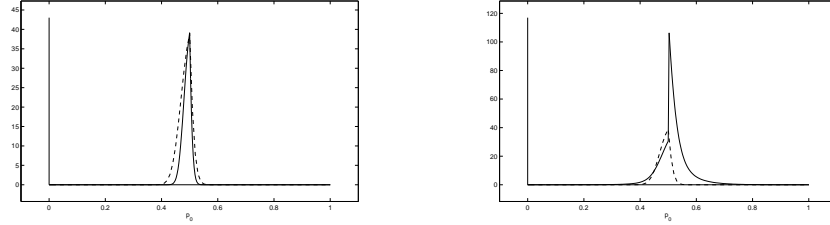


Figure 2: Risk functions $Risk(p_0)$ for the matrix Σ_1 and $\alpha = 0, 1$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axe represents the value of p_0 .

To compare the risk functions we use two type of multivariate distributions (multivariate Gaussian and Student distributions) and three types of correlation matrices:

1. Zero correlations matrix $\Sigma_1 = \text{diag}(1, 1, \dots, 1)$
2. Real correlations matrix Σ_2 calculated from stocks of Dow-Jones index of USA market for 2013.

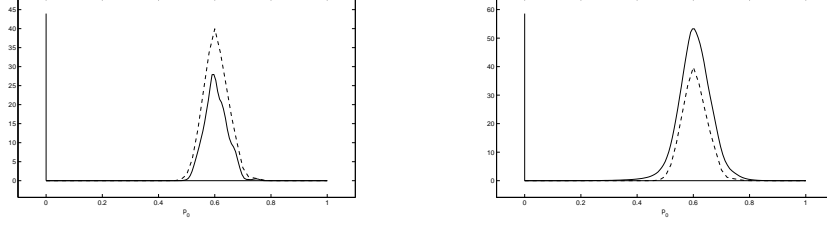


Figure 3: Risk functions $Risk(p_0)$ for the matrix Σ_2 and $\alpha = 0, 5$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axis represents the value of p_0 .

3. High correlations matrix $\Sigma_3 = (\sigma_{i,j})$, with $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0,9$, $i \neq j$.

We choose the following values of parameters: $N = 30$, $n = 400$, significance level for all individual tests $\alpha = 0,5$, $\alpha = 0,1$.

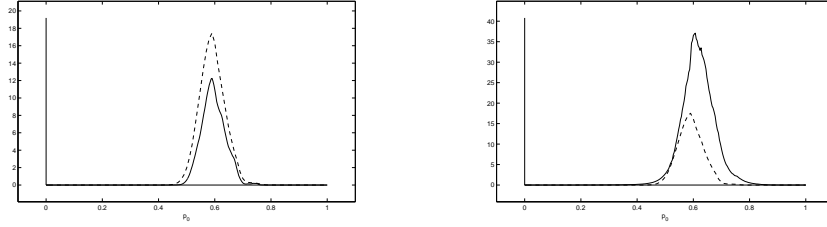


Figure 4: Risk functions $Risk(p_0)$ for the matrix Σ_2 and $\alpha = 0, 1$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axis represents the value of p_0 .

We are interested in behavior of risk function as a function of threshold. For Pearson correlation network the value of threshold ρ_0 is taken from the interval $[-1, 1]$. For sign similarity network the value of threshold p_0 is taken from the interval $[0, 1]$. To make a correct comparison we use the following transformation formula from the Pearson correlation to the probability of sign coincidence:

$$p = \frac{1}{2} + \frac{1}{\pi} \arcsin(\rho)$$

This formula is known for Gaussian distribution ([15], Ch.21), and it can be proved for Student distribution too.

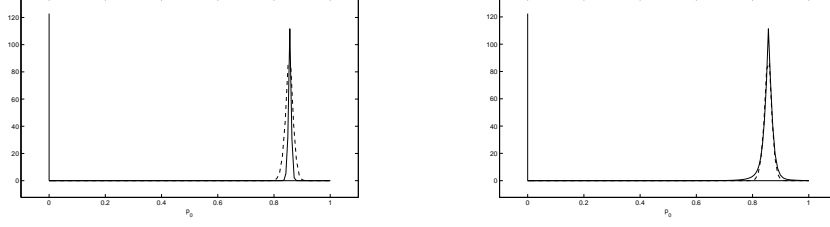


Figure 5: Risk functions $Risk(p_0)$ for the matrix Σ_3 and $\alpha = 0, 5$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axe represents the value of p_0 .

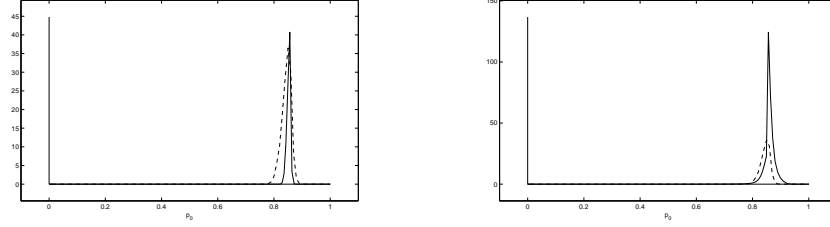


Figure 6: Risk functions $Risk(p_0)$ for the matrix Σ_3 and $\alpha = 0, 1$. Left: Gaussian distribution. Right: Student distribution. Solid line - Pearson correlation network. Dashed line - sign similarity network. Horizontal axe represents the value of p_0 .

The results of numerical experiments are presented in Figures 1-6. Figures 1-2 present the behavior of risk function of δ^{Sg} and δ^P as a function of threshold for the correlation matrix Σ_1 , and $\alpha = 0, 5$ (Figure 1), $\alpha = 0, 1$ (Figure 2). One can see that both procedures control the risk function for Gaussian distribution. Namely, with the change of α from 0,5 to 0,1 the maximal value of risk function for both procedures is decreasing approximately from 120 to 40. In contrast, for Student distribution the maximal value of risk function for the procedure δ^{Sg} is still decreasing, but the maximal value of risk function for the procedure δ^P keeps the same value. This phenomenon is confirmed for the matrix Σ_2 by comparison of Figure 3 and Figure 4 and for the matrix Σ_3 by comparison of Figure 5 and Figure 6. It means that the procedure δ^{Sg} controls the risk function for both distributions while the procedure δ^P does not. This gives advantage to the procedure δ^{Sg} for multivariate Student distributions for small value of α .

8 Concluding remarks

In this paper we introduce and investigate a class of statistical procedures with high reliability for the market graph identification in sign similarity network. Theoretical investigation is conducted in the framework of multiple decision theory. Optimality of multiple decision procedure δ^{Sg} is proved under the following assumptions: additivity of loss functions, unbiasedness of procedures, sign symmetry conditions and known expectations $E(X_i)$, $j = 1, 2, \dots, N$. Additivity of the loss function and unbiasedness of procedures are appropriate for considered problems. Sign symmetry conditions are satisfied for a large class of distributions used in financial analysis, in particular for elliptically contoured distributions. This class includes multivariate Gaussian and Student distributions with heavy tails. Practical advantage of constructed procedures with respect to traditional ones is a high reliability of identification in a larger class of distributions.

Aknowledgement: The first author is supported by RFFI grant 14-01-00807, the second author is supported by RHF grant 15-32-01052, all authors are partly supported by National Research University Higher School of Economics Scientific Found for International research laboratories.

References

- [1] Bautin G.A., Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M. Simple measure of similarity for the market graph construction Computational Management Science, 10, 105-124 (2013).
- [2] Bautin G.A., Kalyagin V. A., Koldanov A. P.: Comparative Analysis of Two Similarity Measures for the Market Graph Construction. Springer Proceedings in Mathematics and Statistics. vol. 59. pp. 29-41 (2013).
- [3] Boginsky V., Butenko S., Pardalos P.M. On structural properties of the market graph. In: Nagurney A. (Editor) Innovations in financial and economic networks. Northampton: Edward Elgar Publishing Inc., 29-45 (2003).
- [4] Boginski V., Butenko S., Pardalos P.M.: Statistical analysis of financial networks. J. Computational Statistics and Data Analysis. 48 (2), 431-443 (2005).
- [5] Boginski V., Butenko S., Pardalos P.M.: Mining market data: a network approach J. Computers and Operations Research. 33 (11) 3171-3184 (2006).
- [6] Boginski V., Butenko S., Pardalos P.M.: Network model of massive data sets, Computer Science and Information Systems, v.1 (2004), pp. 75-89

- [7] Emmert-Streib F. Dehmer M. Identifying critical financial networks of the DJIA: towards a network based index *Complexity*, v.16, n.1, pp. 24-33 (2010).
- [8] Garas A., Argyrakis P., Correlation study of the Athens stock exchange, *Physica A* 380 (2007) 399-410.
- [9] Gunawardena, Athula D. A.; Meyer, Robert R.; Dougan, William L.; Monaghan, Patrick E.; Basu, Choton, Optimal Selection of an Independent Set of Cliques in a Market Graph, *International Proceedings of Economics Development and Research*; Vol. 29, p. 281285 (2012).
- [10] Huang W-Q Zhuang X-T Yao S. A network analysis of the Chinese stock market, *Physica A*, v.388 (2009), pp.2956-2964.
- [11] Hero A., Rajaratnam B.: Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory* 58 (9) 6064 – 6078 (2012).
- [12] Huffner F., Komusiewicz C., Moser H., Niedermeier R. Enumerating Isolated Cliques in Synthetic and Financial Networks, *Combinatorial Optimization and Applications, Lecture Notes in Computer Science*, Volume 5165, pp 405-416 (2008).
- [13] Kenett D.Y., Tumminello M., Madi A., Gur-Gershgoren G., Mantegna R.N., Ben-Jacob E. Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market. *PLoS ONE* 5(12): e15032, 2010. DOI:10.1371/journal.pone.0015032
- [14] Koldanov A.P., Koldanov P.A., Kalyagin V.A., Pardalos P.M. Statistical procedures for the market graph construction.: *Computational Statistics and Data Analysis* 68 17–29 (2013).
- [15] Kramer H. *Mathematical methods of statistics*, Princeton University Press, Ninth Edition 1962.
- [16] Lehmann E.L.: A theory of some multiple decision procedures 1. *Annals of Mathematical Statistics*, 28, 1–25 (1957).
- [17] Lehmann E.L., Romano J.P.: *Testing statistical hypotheses*. Springer, New York, (2005).
- [18] Mantegna R.N., Stanley H.E. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge: Cambridge University Press, (2000).
- [19] Namaki A. Shirazi A.H. Raei R. Jafari G.R. Network analysis of a financial market based on genuine correlation and threshold method, *Physica A*, v.390 (2011), pp. 3835-3841

- [20] Onella J.-P. Kaski K. Kertesz J. Clustering and information in correlation based financial networks, *The European Physical Journal B - Condensed Matter and Complex Systems*, Volume 38 (2004), Issue 2, pp 353-362
- [21] Shirokikh J. Pastukhov G. Boginski V. Butenko S. Computational study of the US stock market evolution: a rank correlation-based network model, *Computational Management Science*, Volume 10, Issue 2-3, pp 81-103 (2013).
- [22] Tse C.K., J. Liu, F.C.M. Lau, A network perspective of the stock market, *Journal of Empirical Finance* 17, 659-667 (2010).
- [23] Tumminello M. Coronello C. Lillo F. Micciche S. Mantegna R. Spanning trees and bootstrap reliability estimation in correlation-based network, *International Journal of Bifurcation and Chaos*, v.17 (2007), pp.2319-2329.
- [24] Vizgunov A. N., Goldengorin B., Kalyagin V. A., Koldanov A. P., Koldanov P., Pardalos P. M. Network approach for the Russian stock market, *Computational Management Science*, v.11, pp.45-55, 2014.
- [25] Wald A.: *Statistical Decision Function*. John Wiley and Sons, New York (1950).
- [26] Wang G-J; Chi X.; Han F. Sun B. Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. *Physica A: Statistical Mechanics and its Applications* 391.16 (2012): 4136-4146.